How to understand the cell by breaking it: network analysis of gene perturbation screens

Florian Markowetz*

Cancer Research UK Cambridge Research Institute, Cambridge CB2 0RE, UK

∗ florian.markowetz@cancer.org.uk

Abstract

Modern high-throughput gene perturbation screens are key technologies at the forefront of genetic research. Combined with rich phenotypic descriptors they enable researchers to observe detailed cellular reactions to experimental perturbations on a genome-wide scale. This review surveys the current state-of-the-art in analyzing single gene perturbation screens from a network point of view. We describe approaches to make the step from the parts list to the wiring diagram by using phenotypes for network inference and integrating them with complementary data sources. The first part of the review describes methods to analyze one- or low-dimensional phenotypes like viability or reporter activity; the second part concentrates on high-dimensional phenotypes showing global changes in cell morphology, transcriptome or proteome.

Introduction

Functional genomics has demonstrated considerable success in inferring the inner working of a cell through analysis of its response to various perturbations. In recent years several technological advances have pushed gene perturbation screens to the forefront of functional genomics. Most importantly, modern technologies make it possible to probe gene function on a genome-wide scale in many model organisms and human. For example, large collections of knock-out mutants play a prominent role in the study of *S. cerevisiae* [1] and RNA interference (RNAi) has become a widely used high-throughput method to knock-down target genes in a wide range of organisms, including *Drosophila melanogaster*, *C. elegans*, and human [2–4].

Another major advance is the development of rich phenotypic descriptions by imaging or measuring molecular features globally. Observed phenotypes can reveal which genes are essential for an organism, or work in a particular pathway, or have a specific cellular function. Combining high-throughput screening techniques with rich phenotypes enables researchers to observe detailed reactions to experimental perturbations on a genome-wide scale. This makes gene perturbation screens one of the most promising tools in functional genomics.

Advances in the design and analysis of gene perturbation screens may have an immediate impact on many areas of biological and medical research. New screening and phenotyping techniques often directly translate into new insights in gene and protein functions. Results of perturbation screens can also reveal unexploited areas of potential therapeutic intervention. For example, a recent RNAi screen showed that some of the most critical protein kinases for the proliferation and survival of cancer cell lines are also the least studied [5].

A goal becoming more and more prominent in both experimental as well as computational research is to leverage gene perturbation screens to the identification of molecular interactions, cellular pathways and regulatory mechanisms. Research focus is shifting from understanding the phenotypes of single proteins to understanding *how* proteins fulfill their function, *what* other proteins they interact with and *where* they act in a pathway. Novel ideas on how to use perturbation screens to uncover cellular wiring diagrams can lead to a better understanding of how cellular networks are de-regulated in diseases like cancer. This knowledge is indispensable for finding new drug targets to attack the drivers of a disease and not only the symptoms.

Phenotypes A phenotype can be any observable characteristic of an organism. Analysis strategies strongly depend on how rich and informative phenotype descriptors are. We will call phenotypes resulting from a single reporter (or a small number of reporters) *low-dimensional* phenotypes and the genes showing significant results *hits* [6, 7]. Examples of such low-dimensional phenotypes are cell viability versus cell death [1], growth rates [8] or the activity of reporter constructs, e.g. a luciferase, downstream of a pathway of interest [9]. Low-dimensional phenotyping screens can identify candidate genes on a genome-wide scale and are often used as a first step for follow-up analysis. We will discuss methods to functionally interpret hits from low-dimensional phenotyping screens and to place them in the context of cellular networks in the first part of this review.

The second part will be devoted to *high-dimensional* phenotyping screens, which evaluate a large number of cellular features at the same time. Observing system-wide changes promises key insights into cellular mechanisms and pathways that can not be supplied by low-dimensional screens. For example, high-dimensional phenotypes can include changes in cell morphology [10–13], or growth rates under a wide range of conditions [14], or transcriptional changes measured on microarrays [15–18], or changes in the metabolome and proteome [19] measured by mass spectrometry [20] or flow cytometry [21, 22]. Morphological and growth phenotypes can be obtained on a genome-wide scale [13, 14], while transcriptional and proteomic phenotypes are often restricted to individual pathways or processes [16, 17, 21].

The distinction between low- and high-dimensional phenotypes may sound technical, but it is crucial for choosing potential analysis methods. The central difference is that high-dimensional phenotypes allow to compute correlations and other similarity measures, which are not applicable for low-dimensional phenotypes. Another important distinction is between *static* phenotypes, providing a 'snapshot' of a cell's reaction to a gene perturbation, and *dynamic* phenotypes showing a cell's reaction over time. We expect more and more studies in the future to produce dynamic output and in the following note explicitly which methods can be applied to dynamic phenotypes. For the biological interpretation of screening results it is very important to keep in mind which level of 'cellular granularity' a phenotype describes: growth rates or cell morphologies are much more 'high-level' features of the cell than gene or protein expressions. As soon as more studies produce dynamic phenotypes on many different cellular levels, integrative analysis of inter-connected phenotypes [23] will become more important. In the following, however, we concentrate on the current state-of-the art, which almost always uses a single type of readout in a perturbation screen.

Pre-processing pipeline In this review we focus on single gene perturbations by knockouts [1] or RNA interference [4] that allow targeting individual genes or combinations of genes. Before network analysis, the raw data needs to pass an initial analysis and quality control pipeline specific to the perturbation and phenotyping technologies used. Low-dimensional screens are mostly performed in multiple-well-plates and a typical analysis pipeline [4] includes data pre-processing, removal of spatial biases per plates, normalization between plates, and finally detection of significant hits [6, 7, 24]. In vertebrates, genes need to be targeted with multiple siRNAs to ensure effective down-regulation [4] and the multiple phenotypes per gene can afterwards be integrated into a statistical score [25]. High-dimensional morphological screens depend on computational analysis like image segmentation [26, 27] and phenotype discovery [28–30] for rapid and consistent phenotyping. Molecular high-dimensional phenotypes need pre-processing depending on their platform and different approaches exist e.g. for flow-cytometry data [31] or microarrays [32].

From phenotypes to cellular networks The phenotypes we have discussed above allow only an indirect view on how different genes in the same process interact to achieve a particular phenotype. Cell morphology or sensitivity to stresses, for example, are global features of the cell and hard to relate directly to how individual genes contribute to them (see Fig. 1a). Gene expression phenotypes show transcriptional changes in the genes downstream of a perturbed pathway but offer only an indirect view of pathway structure due to the high number of non-transcriptional regulatory events like protein modifications [33]. For example, different protein activation states by phosphorylation may not be visible

by changes in mRNA concentrations (see Fig. 1b).

This gap between observed phenotypes and underlying cellular networks is the main problem in the analysis of perturbation screens and applies to both low- and high-dimensional screens. The goal of computational analysis is to bridge this gap by inferring gene function and recovering pathways and mechanism from observed phenotypes. The following methods address the challenge in different ways, mostly by integrating the perturbation effects and phenotypes with additional sources of information like collections of functionally related gene sets or protein-interaction networks.

Network analysis of low-dimensional phenotypes

Global overview by enrichment analysis A simple way to link phenotypes to gene function is to test whether pathways or functional groups of genes (e.g. defined by Gene Ontology terms [34] or MSigDB [35]) are enriched in the list of hits. Most methods use a hypergeometric test statistic (see Fig. 2a) and many can be used online [36–38] or as Bioconductor packages [39]. An alternative global functional annotation method tests whether functional groups show a trend towards especially strong or weak phenotypes without using a cutoff to define hits [35] (see Fig. 2b). Enrichment analysis can also be very useful to analyze high-dimensional phenotypes, for example when functionally annotating the results of a clustering method.

Enrichment analysis results in a list of p-values describing how significantly each gene set was represented in the hits. Enrichment analysis reduces complexity and improves interpretability of results by moving from single genes to functionally related gene sets. This type of analysis is often called 'un-biased' and 'hypothesis-free' and is ideal for a comprehensive first overview. However, enrichment analysis loses its value for complexity reduction if the number of gene sets becomes too big. Also, overlap and dependencies between gene lists that could potentially bias the results have so far only been addressed for the GO graph [38, 39] but not for more general collections of gene lists like MSigDB [35].

Good data analysis asks specific questions. A hypothesis-free method can only be the very first starting point for a deeper exploration of the data. For example, all enrichment methods rely on known gene sets and cannot uncover new pathways or components. Enrichment methods treat pathways as bags of unconnected genes without considering connections within and between pathways. Thus, enrichment methods can only deliver a very crude picture of the cell. In the following we will discuss approaches to overcome some of the limitations of enrichment analysis by integrating the observed phenotypes with complementary sources of information.

Mapping phenotypes to networks Another valuable source of information to interpret RNAi hits are gene and protein networks obtained either experimentally [40, 41] or computationally by literature mining [42] or integrating heterogeneous genomic data [43–45]. All computational networks are available online on supplementary webpages and the experimental networks can be obtained from databases like STRING [46] or BioGRID [47].

Using these complementary data sources can improve hit identification [48–50] and even provide a more refined view of the pathways the hits contribute to. One strategy is to search for sub-networks containing a surprisingly large number of hits (see Fig. 3a). While this strategy is already useful when evaluating interesting sub-networks by eye [51, 52] its true power comes from the availability of efficient search algorithms to find sub-networks enriched for RNAi hits and assess their significance [53–57]. An additional application of mapping hits to a network is that known phenotypes can be used to predict phenotypes of genes not included in the screen, e.g. by assuming that a gene connected to many hits should also show a strong phenotype [51]. The success of all network-mapping strategies strongly depends on the quality and coverage of both the screen and the linkage in the network.

Gene prioritization Other approaches complement genomic data with biological prior knowledge showing how 'interesting' hits look like. Gene prioritization [49,58] ranks genes according to how promising they would be for follow-up studies. Because it uses prior knowledge to fine-tune the algorithm, gene prioritization can be more focussed than a global un-informed search for enriched subnetworks.

Network analysis of high-dimensional phenotypes

Global overview by clustering and ranking. Most state-of-the-art analysis techniques rely on a guilt-by-association paradigm: genes with similar phenotypes will most probably have a similar biological function. This explains the prevalence of clustering techniques in analyzing high-dimensional phenotyping screens [10,13,14,17]. Clustering is a convenient first analysis and visualization step that can can highlight strong trends and patterns in the data and can thus yield a global first impression of functional units. Another analysis strategy relying on guilt-by-association is to rank genes by their phenotypic similarity compared to a gene of interest [11]. Clustering and ranking can be combined with enrichment analysis (as discussed above) for functional interpretation.

Graph methods linking causes to effects Another useful data visualization especially for transcriptional phenotypes is to build a directed (not necessarily acyclic) graph by drawing an arrow between two genes if perturbing one results in a significant expression change at the other [59]. This graph representation can be then used as a starting point for further analysis, for example by using graph-theoretic methods of transitive reduction [60] to distinguish between direct and indirect effects of a perturbation [61,62].

Probabilistic graphical models. Most approaches to infer pathway structure from experimental data rely on probabilistic graphical models. For low-dimensional phenotypes they often suffer from non-uniqueness and un-identifiability issues [63], but can be applied very successfully in high-dimensional settings. A prominent approach are (static or dynamic) Bayesian networks, which describe probabilistically how a gene is controlled by its regulators [64,65]. To model experimental perturbations most approaches rely on the concept of 'ideal interventions' [66] which deterministically fix a target gene to a particular state (e.g '0' for a gene knockout). Ideal interventions were applied in Bayesian networks [21,67,68], factor graphs [69] and dependency networks [70]. In simulations [71,72] and on real data [21,71] it was found that interventions are critical for effective inference.

The model of ideal interventions contains a number of idealizations (hence the name), most importantly that manipulations only affect single genes and that perturbation strength can be controlled deterministically. The first assumption may not be true if there are off-target or compensatory effects involving other genes. The second assumption may also not hold true in realistic biological scenarios; in particular for RNAi screens where experimentalists often lack knowledge about the exact knock-down efficiency. Probabilistic generalizations of ideal interventions can be used to cope with this uncertainty [73].

Probabilistic data integration High-dimensional phenotypic profiles can be mapped to given graphs and networks by finding subgraphs that are connected in the background network and at the same time show high similarity of phenotypic profiles. These approaches already exist for mapping gene expression data onto protein interaction networks [74] and the same algorithms could easily be applied to any other kind of high-dimensional phenotypic profiles (see Fig. 3b). Other approaches use data integration to construct potential pathways from protein interactions and transcription factor binding data to relate perturbed genes to the observed downstream effects [75–77].

Multiple Input - Multiple Output (MIMO) models Many of the approaches discusses so far–like clustering or graphical models–can be applied to both static 'snapshots' as well as dynamic time-course measurements. Another approach to model specifically the dynamics of networks comes from a

branch of control theory called 'systems identification' [78] and uses so called Multiple Input - Multiple Output (MIMO) models. MIMO models represent the evolution of a perturbed cell over time by linear differential equations [79–83] and can represent non-linear effects by transfer functions [84]. The models can be inferred by regression techniques in the linear case [80] or Monte Carlo stochastic search in the non-linear case [84]. The framework is very flexible and can incorporated single as well as combinatorial perturbations.

Nested Effects Models (NEMs) One of the key problems in analyzing perturbation screens is that the observed phenotypes are downstream of the perturbed pathway and may not show the direct influence of one pathway component on another. A class of models explicitly addressing this problem are Nested Effects Models [33, 85]. They reconstruct pathway structure from *subset relations* based on the following rationale: Perturbing some genes may have an influence on a global process, while perturbing others affects sub-processes of it. Imagine, for example, a signaling pathway activating several transcription factors. Blocking the entire pathway will most probably affect all targets of all transcription factors, while perturbing a single transcription factor will only affect its direct targets, which are a subset of the phenotype obtained by blocking the complete pathway. Given high-dimensional phenotypes showing a subset structure, NEMs find the most likely pathway topology explaining the data. They differ from other statistical approaches like Bayesian networks by encoding subset relations instead of correlations or other similarity measures. The theory of NEMs has been applied and extended in several studies [86–89]. An implementation is available as an R/Bioconductor package [90]. Other extensions to the NEM framework distinguish between activating and inhibiting regulation [91] or include dynamic information from time-series measurements [92].

Discussion and Outlook

In this review we have discussed two main approaches to describe the reaction of a cell to an experimental gene perturbation: low-dimensional phenotypes measure individual reporters for cell viability or pathway activation, while high-dimensional phenotypes show global effects on cell morphology, transcriptome or proteome. Table 1 lists examples of freely available software implementing some of these approaches. All of them can be directly applied to gene perturbation screens, even though some of them have been introduced in different contexts. While this review has focused on single gene knock-outs and knock-downs, similar approaches can be applied to gene over-expression screens [22, 83, 93, 94], drug treatment [84], environmental stresses changing many genes [95, 96] or even natural genetic variation [97].

Predicting phenotypes from metabolic networks The focus of this review is on functionally annotating hits in a network context and reconstructing networks from high-dimensional phenotypes. In a complementary direction of research, genome-wide reconstructions of metabolic networks [98, 99] are used to predict effects of gene perturbations. Instead of predicting networks from phenotypes, these approaches predict phenotypes from networks. For example, in *S. cerevisiae* and *E. coli* computational models very accurately predict fitness effects of gene knock-outs [100, 101] as well as compensatory rescue effects [102]. However, recent developments in metabolic network modeling have led to linear programming algorithms to extract relevant context-specific sub-networks of activity from a genome-wide network [103, 104]. In the same way as the probabilistic data integration methods discussed above, e.g. [74], these algorithms could be used in the future to find metabolic sub-networks active under certain gene perturbations.

From single to combinatorial perturbations While single gene perturbation screens have been immensely successful in extending our knowledge of pathway components and interactions, an important limitation can be caused by compensatory effects, genetic buffering and redundancy of cellular mechanisms and pathways [105, 106]. This can only be overcome by perturbing several genes at the same time.

The number of possible combinations grows rapidly and thus current approaches are mainly limited to perturbing pairs of genes and observing low-dimensional phenotypes like fitness estimates [107]. The analysis of combinatorial perturbations is the topic of another review [108].

The end of the screen is the beginning of the experiment Global phenotyping and pathway screening can be combined in the same study. For example, a first genome-wide screen identifies key genes representative for pathways and cellular mechanisms involved in the phenotype. In a second step the hits of the first screen could be assayed for high-dimensional molecular phenotypes to infer a pathway diagram using Nested Effects Models or other statistical approaches.

In a further step this preliminary pathway models could be used to plan an additional round of experimentation. Different modeling frameworks propose future experiments to most effectively refine a pathway hypothesis, e.g. Bayesian networks [109,110], physical network models [76], logical models [111], Boolean networks [112], and dynamical modeling [79].

Iteratively integrating experimentation and computation may lead to a virtuous circle and is one of the most promising approaches to refine our understanding of the inner working of the cell.

Acknowledgments

References

1. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, et al. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285: 901–906.

2. Fuchs F, Boutros M (2006) Cellular phenotyping by RNAi. Brief Funct Genomic Proteomic 5: 52–56.

3. Moffat J, Sabatini D (2006) Building mammalian signalling pathways with RNAi screens. Nature Reviews Molecular Cell Biology 7: 177–187.

4. Boutros M, Ahringer J (2008) The art and design of genetic screens: RNA interference. Nat Rev Genet 9: 554–566.

5. Luo B, Cheung H, Subramanian A, Sharifnia T, Okamoto M, et al. (2008) Highly parallel identification of essential genes in cancer cells. Proceedings of the National Academy of Sciences of the United States of America 105: 20380–20385.

6. Boutros M, Brás LP, Huber W (2006) Analysis of cell-based RNAi screens. Genome Biol 7: R66.

7. Rieber N, Knapp B, Eils R, Kaderali L (2009) RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens. Bioinformatics .

8. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the saccharomyces cerevisiae genome. Nature 418: 387–391.

9. Müller P, Kuttenkeuler D, Gesellchen V, Zeidler MP, Boutros M (2005) Identification of JAK/STAT signalling components by genome-wide RNA interference. Nature 436: 871–875.

10. Perlman ZE, Slack MD, Feng Y, Mitchison TJ, Wu LF, et al. (2004) Multidimensional drug profiling by automated microscopy. Science 306: 1194–1198.

11. Gunsalus KC, Yueh WC, MacMenamin P, Piano F (2004) RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. Nucleic acids research 32: D406–10.

12. Neumann B, Held M, Liebel U, Erfle H, Rogers P, et al. (2006) High-throughput RNAi screening by time-lapse imaging of live human cells. Nature Methods 3: 385–390.

13. Bakal C, Aach J, Church G, Perrimon N (2007) Quantitative morphological signatures define local signaling networks regulating cell morphology. Science 316: 1753–1756.

14. Brown JA, Sherlock G, Myers CL, Burrows NM, Deng C, et al. (2006) Global analysis of gene function in yeast by quantitative phenotypic profiling. Mol Syst Biol 2: 2006.0001.

15. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional discovery via a compendium of expression profiles. Cell 102: 109–126.

16. Boutros M, Agaisse H, Perrimon N (2002) Sequential activation of signaling pathways during innate immune responses in drosophila. Dev Cell 3: 711–722.

17. Ivanova N, Dobrin R, Lu R, Kotenko I, Levorse J, et al. (2006) Dissecting self-renewal in stem cells with RNA interference. Nature 442: 533–533.

18. Amit I, Garber M, Chevrier N, Leite AP, Donner Y, et al. (2009) Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. Science .

19. Ideker T, Thorsson V, Ranish J, Christmas R, Buhler J, et al. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science 292: 929–934.

20. Gstaiger M, Aebersold R (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. Nat Rev Genet 10: 617–627.

21. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308: 523–529.

22. Niu W, Li Z, Zhan W, Iyer VR, Marcotte EM (2008) Mechanisms of cell cycle control revealed by a systematic and quantitative overexpression screen in s. cerevisiae. PLoS Genetics 4: e1000120.

23. Lu R, Markowetz F, Unwin R, Leek J, Airoldi E, et al. (2009) Systems-level dynamic analyses of fate change in murine embryonic stem cells. Nature 462(7271):358-362

24. Birmingham A, Selfors L, Forster T, Wrobel D, Kennedy C, et al. (2009) Statistical methods for analysis of high-throughput RNA interference screens. Nature Methods 6: 569–575.

25. König R, Chiang Cy, Tu B, Yan SF, DeJesus P, et al. (2007) A probability-based approach for the analysis of large-scale RNAi screens. Nature Methods advanced online publication: 847–849.

26. Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, et al. (2006) Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology 7: R100.

27. Sklyar O, Pau G, Smith M, Huber W (2008) EBimage: Image Processing and Image Analysis Toolkit for R.

28. Jones TR, Carpenter AE, Lamprecht MR, Moffat J, Silver SJ, et al. (2009) Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. Proc Natl Acad Sci U S A 106: 1826–1831.

29. Yin Z, Zhou X, Bakal C, Li F, Sun Y, et al. (2008) Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput rnai screens. BMC Bioinformatics 9: 264.

30. Wang J, Zhou X, Bradley PL, Chang SF, Perrimon N, et al. (2008) Cellular phenotype recognition for high-content rna interference genome-wide screening. J Biomol Screen 13: 29–39.

31. Hahne F, Arlt D, Sauermann M, Majety M, Poustka A, et al. (2006) Statistical methods and software for the analysis of high throughput reverse genetic assays using flow cytometry readouts. Genome Biology 7: R77.

32. Smyth GK (2005) Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, R Irizarry WH, editors, Bioinformatics and Computational Biology Solutions using R and Bioconductor, New York: Springer. pp. 397–420.

33. Markowetz F, Bloch J, Spang R (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. Bioinformatics 21: 4026–4032.

34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. Nat Genet 25: 25–29.

35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550.

36. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc 4: 44–57.

37. Sealfon RSG, Hibbs MA, Huttenhower C, Myers CL, Troyanskaya OG (2006) GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. BMC Bioinformatics 7: 443.

38. Bauer S, Grossmann S, Vingron M, Robinson PN (2008) Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. Bioinformatics 24: 1650–1651.

39. Alexa A, Rahnenführer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22: 1600–1607.

40. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, et al. (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14: 292–299.

41. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968.

42. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, et al. (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. Science 309: 1078–1083.

43. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306: 1555–1558.

44. Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, et al. (2005) Discovery of biological networks from diverse functional genomic data. Genome Biol 6: R114.

45. Guan Y, Myers CL, Lu R, Lemischka IR, Bult CJ, et al. (2008) A genomewide functional network for the laboratory mouse. PLoS Comput Biol 4: e1000165.

46. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) String 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–D416.

47. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The BioGRID interaction database: 2008 update. Nucleic Acids Res 36: D637–D640.

48. Kaplow I, Singh R, Friedman A, Bakal C, Perrimon N, et al. (2009) RNAiCut: automated detection of significant genes from functional genomic screens. Nature Methods 6: 476–477.

49. Wang L, Tu Z, Sun F (2009) A network-based integrative approach to prioritize reliable hits from multiple genome-wide rnai screens in drosophila. BMC Genomics 10: 220.

50. Berndt JD, Biechele TL, Moon RT, Major MB (2009) Integrative analysis of genome-wide RNA interference screens. Science Signaling 2: pt4.

51. Lee I, Lehner B, Crombie C, Wong W, Fraser A, et al. (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in caenorhabditis elegans. Nature Genetics 40: 181–188.

52. Krishnan M, Ng A, Sukumaran B, Gilfoy F, Uchil P, et al. (2008) RNA interference screen for human genes associated with west nile virus infection. Nature 455: 242–245.

53. Ideker T, Ozier O, Schwikowski B, Siegel AF (2002) Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1: S233–S240.

54. König R, Zhou Y, Elleder D, Diamond TL, Bonamy GMC, et al. (2008) Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell 135: 49–60.

55. Dittrich M, Klau G, Rosenwald A, Dandekar T, Müller T (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24: i223.

56. Bankhead A, Sach I, Ni C, LeMeur N, Kruger M, et al. (2009) Knowledge based identification of essential signaling from genome-scale siRNA experiments. BMC Systems Biology 3: 80.

57. Tu ZCA, Wong K, Mitnaul L, Edwards S, Sach I, et al. (2009) Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network. Genome Research 19: 1057.

58. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotech 24: 537–544.

59. Rung J, Schlitt T, Brazma A, Freivalds K, Vilo J (2002) Building and analysing genome-wide gene disruption networks. Bioinformatics 18: 202–210.

60. Aho AV, Garey MR, Ullman JD (1972) The transitive reduction of a directed graph. SIAM Journal on Computing 1: 131-137.

61. Wagner A (2001) How to reconstruct a large genetic network from n gene perturbations in fewer than n^2 easy steps. Bioinformatics (Oxford, England) 17: 1183–1197.

62. Tresch A, Beissbarth T, Sültmann H, Kuner R, Poustka A, et al. (2007) Discrimination of direct and indirect interactions in a network of regulatory effects. Journal of computational biology : a journal of computational molecular cell biology 14: 1217–1228.

63. Kaderali L, Dazert E, Zeuge U, Frese M, Bartenschlager R (2009) Reconstructing signaling pathways from rnai data using probabilistic boolean threshold networks. Bioinformatics 25: 2229–2235.

64. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science 303: 799–805.

65. Markowetz F, Spang R (2007) Inferring cellular networks–a review. BMC Bioinformatics 8 Suppl 6: S5.

66. Pearl J (2000) Causality: Models, Reasoning and Inference. Cambridge University Press.

67. Ellis B, Wong WH (2008) Learning causal bayesian network structures from experimental data. Journal of the American Statistical Association 103: 778-789.

68. Peer D, Regev A, Elidan G, Friedman N (2001) Inferring subnetworks from perturbed expression profiles. Bioinformatics 17: 215–224.

69. Gat-Viks I, Tanay A, Raijman D, Shamir R (2006) A probabilistic methodology for integrating knowledge and experiments on biological networks. J Comput Biol 13: 165–181.

70. Rogers S, Girolami M (2005) A Bayesian regression approach to the inference of regulatory networks from gene expression data. Bioinformatics 21: 3131–3137.

71. Werhli AV, Grzegorczyk M, Husmeier D (2006) Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. Bioinformatics 22: 2523–2531.

72. Zak DE, Gonye GE, Schwaber JS, Doyle FJ (2003) Importance of input perturbations and stochastic gene expression in the reverse engineering of genetic regulatory networks: insights from an identifiability analysis of an in silico network. Genome Res 13: 2396–2405.

73. Markowetz F, Grossmann S, Spang R (2005) Probabilistic soft interventions in conditional gaussian networks. Conference on Artificial Intelligence and Statistics (AISTATS) .

74. Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 1: 8.

75. Yeang CH, Ideker T, Jaakkola T (2004) Physical network models. J Comput Biol 11: 243–262.

76. Yeang CH, Mak HC, McCuine S, Workman C, Jaakkola T, et al. (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. Genome biology 6: R62.

77. Ourfali O, Shlomi T, Ideker T, Ruppin E, Sharan R (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. Bioinformatics 23: i359–66.

78. Ljung L (1986) System Identification: Theory for the User. Prentice Hall.

79. Tegner J, Yeung MKS, Hasty J, Collins JJ (2003) Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proceedings of the National Academy of Sciences 100: 5944–5949.

80. Gardner TS, di Bernardo D, Lorenz D, Collins JJ (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301: 102–105.

81. Xiong M, Li J, Fang X (2004) Identification of genetic networks. Genetics 166: 1037–1052.

82. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, et al. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nature Biotechnology 23: 377–383.

83. Lorenz DR, Cantor CR, Collins JJ (2009) A network biology approach to aging in yeast. Proceedings of the National Academy of Sciences of the United States of America 106: 1145.

84. Nelander S, Wang W, Nilsson B, She QB, Pratilas C, et al. (2008) Models from experiments: combinatorial drug perturbations of cancer cells. Molecular systems biology 4: 216.

85. Markowetz F, Kostka D, Troyanskaya O, Spang R (2007) Nested effects models for high-dimensional phenotyping screens. Bioinformatics 23: 305–312.

86. Fröhlich H, Fellmann M, Sueltmann H, Poustka A, Beissbarth T (2007) Large scale statistical inference of signaling pathways from RNAi and microarray data. BMC Bioinformatics 8: 386.

87. Tresch A, Markowetz F (2008) Structure learning in nested effects models. Stat Appl Genet Mol Biol 7.

88. Fröhlich H, Fellmann M, Sültmann H, Poustka A, Beissbarth T (2008) Estimating large scale signaling networks through nested effect models with intervention effects from microarray data. Bioinformatics .

89. Fröhlich H, Tresch A, Beißbarth T (2009) Nested effects models for learning signaling networks from perturbation data. Biometrical journal 51: 304–323.

90. Fröhlich H, Beißbarth T, Tresch A, Kostka D, Jacob J, et al. (2008) Analyzing gene perturbation screens with nested effects models in R and Bioconductor. Bioinformatics 24: 2549–2550.

91. Vaske CJ, House C, Luu T, Frank B, Yeang CH, et al. (2009) A factor graph nested effects model to identify networks from genetic perturbations. PLoS computational biology 5: e1000274.

92. Anchang B, Sadeh M, Jacob J, Tresch A, Vlad M, et al. (2009) Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. Proceedings of the National Academy of Sciences 106: 6447.

93. Sopko R, Huang D, Preston N, Chua G, Papp B, et al. (2006) Mapping pathways and phenotypes by systematic gene overexpression. Molecular cell 21: 319–330.

94. Stokic D, Hanel R, Thurner S (2009) A fast and efficient gene-network reconstruction method from multiple over-expression experiments. BMC Bioinformatics 10: 253.

95. Yosef N, Kaufman A, Ruppin E (2006) Inferring functional pathways from multi-perturbation data. Bioinformatics 22: e539.

96. MacCarthy T, Pomiankowski A, Seymour R (2005) Using large-scale perturbations in gene network reconstruction. BMC Bioinformatics 6: 11.

97. Rockman M (2008) Reverse engineering the genotype-phenotype map with natural genetic variation. Nature 456: 738–744.

98. Herrgard MJ, Swainston N, Dobson P, Dunn WB, Arga KY, et al. (2008) A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nat Biotechnol 26: 1155–1160.

99. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, et al. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci U S A 104: 1777–1782.

100. Papp B, Pl C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429: 661–664.

101. Fong SS, Palsson B (2004) Metabolic gene-deletion strains of escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet 36: 1056–1058.

102. Motter A, Gulbahce N, Almaas E, Barabasi AL (2008) Predicting synthetic rescues in metabolic networks. Mol Sys Bio 4: 168.

103. Shlomi T, Cabili MN, Herrgrd MJ, Palsson B, Ruppin E (2008) Network-based prediction of human tissue-specific metabolism. Nat Biotechnol 26: 1003–1010.

104. Becker SA, Palsson BO (2008) Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol 4: e1000082.

105. Deutscher D, Meilijson I, Schuster S, Ruppin E (2008) Can single knockouts accurately single out gene functions? BMC Systems Biology 2: 50.

106. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, et al. (2009) Backup in gene regulatory networks explains differences between binding and knockout results. Mol Syst Biol 5.

107. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, et al. (2004) Global mapping of the yeast genetic interaction network. Science 303: 808–813.

108. Myers CL (2009) Analysis of combinatorial gene perturbation screens. PLoS Comput Biol IN PREPARATION: (the 'partner review' to this one - what's the best way to cite it?).

109. Pournara I, Wernisch L (2004) Reconstruction of gene networks using Bayesian learning and manipulation experiments. Bioinformatics 20: 2934–2942.

110. Yoo C, Cooper GF (2004) An evaluation of a system that recommends microarray experiments to perform to discover gene-regulation pathways. Artif Intell Med 31: 169–182.

111. Szczurek E, Gat-Viks I, Tiuryn J, Vingron M (2009) Elucidating regulatory mechanisms downstream of a signaling pathway using informative experiments. Mol Syst Biol 5: 287.

112. Ideker TE, Thorsson V, Karp RM (2000) Discovery of regulatory interactions through perturbation: inference and experimental design. Pac Symp Biocomput : 305–316.

113. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5: R80.

114. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
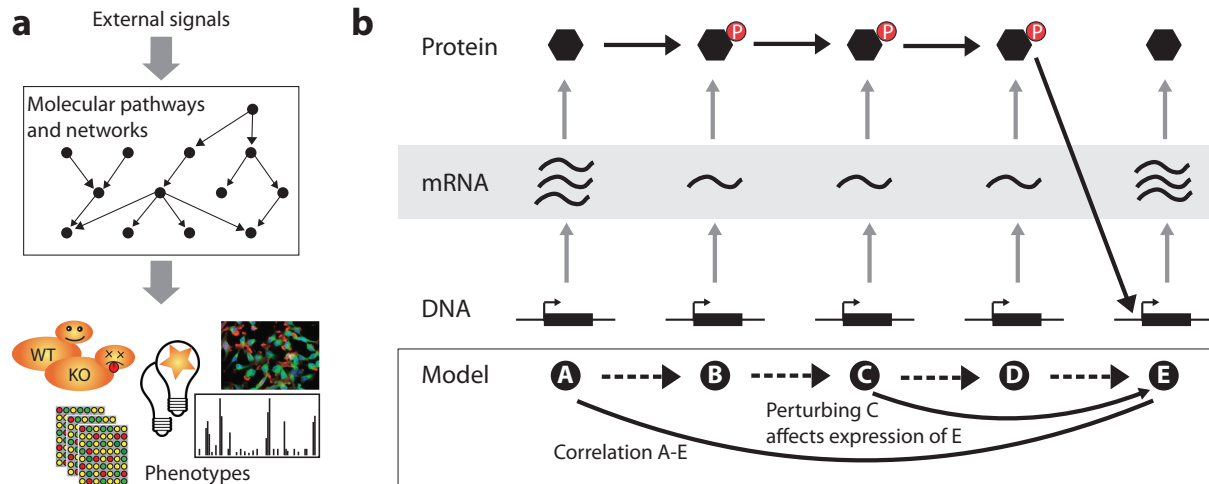
Figures and legends

Figure 1. Cellular networks underlying observable phenotypes. (a) Phenotypes are the response of the cell to external signals mediated by cellular networks and pathways. The goal of computation is to reconstruct these networks from the observed phenotypes. **(b)** Global molecular phenotypes like gene expression allow a view inside the cell but also have limitations. This is exemplified here in a cartoon pathway adapted from [61] showing a cascade of five genes/proteins (A-E). Proteins A-C form a kinase cascade, D is a transcription factor acting on E. Up-regulation of A starts information flow in the cascade and results in E being turned on. In gene expression data this is visible as a correlation between A and E (represented as an undirected edge in the model). Experimentally perturbing a genes, say C, removes the corresponding protein from the cascade, breaks the information flow and results in an expression change at E (represented as an arrow in the model). However, the different phosphorylation and activation states of proteins B-D will most probably not be visible as changes in gene expression. Thus, due to the pathway mostly acting on the protein level most parts of the cascade (dashed arrows in the model) can not be inferred from gene expression data directly.
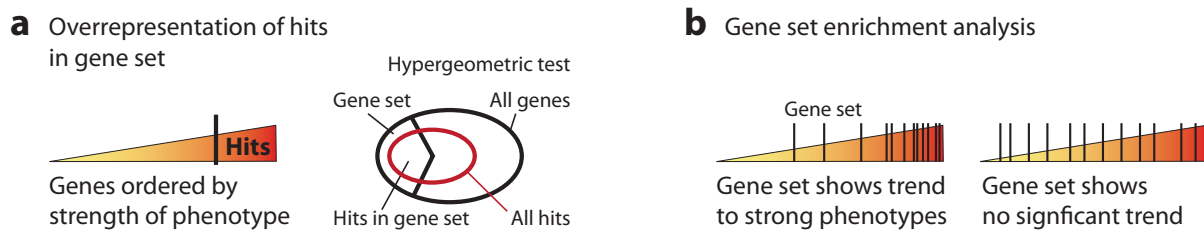

Figure 2. Functional annotation of hits by enrichment analysis. (a) In the first approach [38] a cutoff is applied to select the hits with strongest phenotypes. A hyper-geometric test then evaluates if the overlap between the hits and a given gene set is surprisingly large (or small) compared to the overlap with a random set. **(b)** A second approach [35] does not need a cutoff. It maps the gene set (black bars) onto the observed phenotypes and quantifies if there is a significant trend or if the genes are spread out uniformly over the whole range.
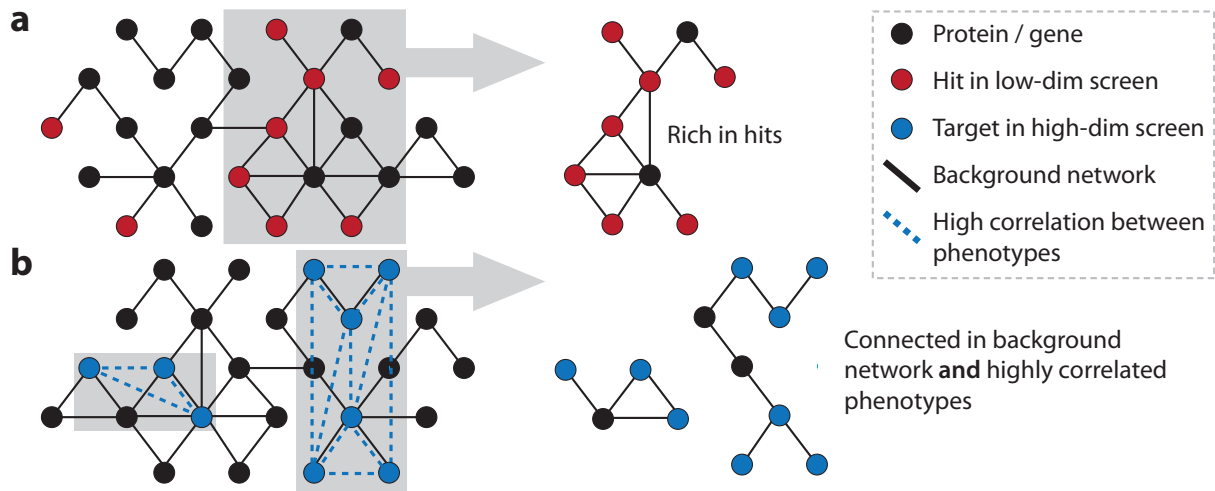
Figure 3. Extracting rich sub-networks. Different patterns in the graph point to a common cellular mechanism causing a phenotype: **(a)** hits in a low-dimensional screen (red nodes) clustering in highly connected sub-networks, and **(b)** high correlation between high-dimensional phenotypes of target genes connected in the background network. The black graph represents any type of background network.

Tables

Table 1. Examples of software for network analysis of gene perturbation screens.

General data analysis and network visualization		
Bioconductor	Software environment for the analysis of genomic data featuring hundreds of contributed packages [113]	www.bioconductor.org
Cytoscape	Software platform for visualizing molecular interaction networks and integrating them with other data types [114]	www.cytoscape.org
Setting up data for network analysis		
cellHTS2	End-to-end analysis of cell-based screens: from raw intensity readings to the annotated hit list [6]	www.bioconductor.org
RNAither	Analysis of cell-based RNAi screens, includes quality assessment and customizable normalization [7]	www.bioconductor.org
EBImage	Cell image analysis and feature extraction [27]	www.bioconductor.org
CellProfiler	Cell image analysis and feature extraction [26]	www.cellprofiler.org
Enrichment analysis		
DAVID	Tools for data annotation, visualization and integration [36]	david.abcc.ncifcrf.gov
GOLEM	Enrichment analysis and visualization of GO graph (Fig 2a) [37]	function.princeton.edu/GOLEM
Ontologizer	Enrichment analysis with dependencies between GO nodes (Fig 2a) [38]	compbio.charite.de/ontologizer
GSEA	Gene set enrichment analysis (Fig 2b) [35]	www.broadinstitute.org/gsea/
Clustering and ranking		
Cell Profiler Analyst	Interactive exploration and analysis of multidimensional data from image-based experiments [28]	www.cellprofiler.org
PhenoBlast	Ranking of phenotype profiles according to similarity with given profile [11]	www.rnai.org
Endeavour	Prioritizes hits for further analysis [58]	www.esat.kuleuven.be/endeavour/
Finding rich sub-networks		
heinz	Finds optimal subnetworks rich in hits (Fig 3a) [55]	www.planet-lisa.net
jActiveModules	Finds heuristic subnetworks rich in hits (Fig 3a) [53]	www.cytoscape.org
Matisse	Finds subnetworks with high phenotypic similarity (Fig 3b) [74]	acgt.cs.tau.ac.il/matisse/
Network reconstruction		
nem	Nested Effects Models reconstruct pathway features from subset relations in high-dim phenotypes [90]	www.bioconductor.org
copia	Copia uses MIMO models to reconstruct networks from perturbations [84]	cbio.mskcc.org/copia/

The table contains the name of the method, a short description with reference, and a webpage where it can be obtained. This list is far from comprehensive, but hopefully provides a starting point even for non-coding experimentalists.